

Active Sampler: Light-weight Accelerator for Complex Data Analytics at Scale

Jinyang Gao[‡] H. V. Jagadish[§] Beng Chin Ooi[‡]

[‡] National University of Singapore [§] University of Michigan

[‡] {jinyang.gao, ooi@comp.nus.edu.sg} [§] jag@umich.edu

ABSTRACT

Recent years have witnessed amazing outcomes from “Big Models” trained by “Big Data”. Most popular algorithms for model training are iterative. Due to the surging volumes of data, we can usually afford to process only a fraction of the training data in each iteration. Typically, the data are either uniformly sampled or sequentially accessed.

In this paper, we study how the data access pattern can affect model training. We propose an *Active Sampler* algorithm, where training data with more “learning value” to the model are sampled more frequently. The goal is to focus training effort on valuable instances near the classification boundaries, rather than evident cases, noisy data or outliers. We show the correctness and optimality of Active Sampler in theory, and then develop a light-weight vectorized implementation. Active Sampler is orthogonal to most approaches optimizing the efficiency of large-scale data analytics, and can be applied to most analytics models trained by stochastic gradient descent (SGD) algorithm. Extensive experimental evaluations demonstrate that Active Sampler can speed up the training procedure of SVM, feature selection and deep learning, for comparable training quality by 1.6-2.2x.

1. INTRODUCTION

We live in an age of ever-increasing size and complexity of “Big Data”. To understand the data and decipher the information that truly counts, many advanced large-scale machine learning models have been devised, from million-dimension linear models (e.g. Logistic Regression [9], Support Vector Machine [25], feature selection [16, 31], Principal Component Analysis [10]) to complex models like Deep Neural Networks [4] or topic models [8]. While these models have demonstrated value for a wide spectrum of applications [18, 16, 12], their complexity causes the training cost to increase dramatically with the surging volume of data. This difficulty with scale severely affects the viability of many advanced models on industry-scale applications. Consequently, accelerating the training procedure of those “Big Models” on “Big Data” has attracted a great deal of interest.

All the models mentioned in the preceding paragraph, and many others, can be formulated as minimizing a specific objective function based on a set of data observations, i.e. the Empirical Risk Minimization [26] (ERM) problem. Even though gradient descent [3, 1] has been widely used for decades, evaluating the full gradient over all the training samples (i.e. batch gradient descent [29]) is extremely

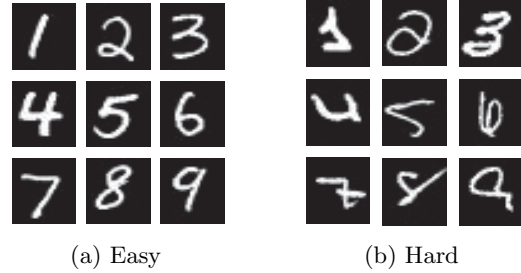


Figure 1: Information Gain from Training Data

expensive in the prevailing scale of millions of training samples. To reduce the computational cost at every iteration, Stochastic Gradient Descent (SGD) [22, 2] optimizes the objective function based on a single random sample at each iteration. Thereby, the computation cost per iteration is reduced greatly, but now many more iterations are required to reach a certain degree of accuracy or finally converge [20]. This is because the stochastic gradient used in each iteration is highly sensitive to the specific random sample chosen. Although the expectation of stochastic gradient is exactly the full gradient, the large variance causes the direction of stochastic gradient to deviate from that of the full gradient, which is the optimal direction to minimize the objective function. Some samples may even direct the model to the opposite of the correct direction.

It has been shown that reducing the variance of stochastic gradient [27, 11, 23] will lead to a much faster convergence rate. A commonly used scheme called Mini-batch SGD [15] is developed for this purpose: by averaging the gradient from a mini-batch of samples, the variance of gradient is significantly reduced, at the cost of some increased computation per iteration. The optimal mini-batch size is determined as a trade-off between the increased computational cost per iteration for a larger mini-batch and the increased variance (and hence number of iterations) for a smaller mini-batch.

In this paper, we seek to further optimize the stochastic gradient by not merely averaging gradients from more random samples but rather improving the quality of data samples. To this end, we propose a light-weight SGD accelerator inspired by active learning [24, 21]. In active learning, training data are selected to maximize the “learning value”. For example, to train a classification model, training data points near class boundaries are more valuable than points in the interior of a class. We adapt the idea of active learning to the SGD optimization context by choosing samples from not

a uniform distribution over the training data but rather a biased distribution from which we expect to learn more.

Figure 1 gives an example of how different samples can affect the training efficiency. The left side contains some images of written digits that are very easy to classify. For these “easy-to-classify” images, most models can classify them correctly even after a handful of steps. In subsequent thousands of iterations or even more, these easy-to-classify images will be sampled and correctly classified with high confidence, contributing almost zero gradient to the model. Consequently, the training time consumed by those easy-to-classify images is largely wasted. In contrast, the right side of Figure 1 contains some images that are hard to classify. By putting more effort on those “hard-to-classify” images, the accuracy of model may improve at a faster rate.

Based on the above intuition, we develop a weighted sampling method called *Active Sampler*. We find that to maximize the information gain in each iteration, the sample frequency for each training sample should be proportional to the estimate of its gradient magnitude. Notwithstanding the sampler itself is biased, we show that the original objective function based on uniform weight can still be correctly minimized by re-weighting the gradient of each sample. From the view of variance reduction, Active Sampler also provides a gradient with the smallest variance compared with all weighted sampling methods, including the uniform random sampling. The net result is a system that requires far fewer iterations for model convergence (or to reach a required accuracy threshold).

Although optimization methods can reduce the number of iterations, it should also be noted that they may introduce additional computation cost in each iteration. There have been a great amount of research works [23, 11, 19] focusing on accelerating SGD. Theoretically, these methods have significantly faster convergence rate. While Momentum and AdaGrad [6] methods have shown their effectiveness and have been integrated into practical SGD solver, most methods are far from being used in practice due to their significant additional computation cost. For example, the cost of SVRG [11] per iteration is at least three times the cost of standard SGD per iteration. As noted in [19], mini-batch SGD still dominates in most cases, due to its light-weight computation and good vectorization.

To make Active Sampler as efficient as possible in practice, the design principle is to reduce the overhead involved in each iteration. In the actual implementation, we use lots of existing knowledge to approximate the information that needs additional computation cost, and the sampling distribution is decided by the gradient magnitude in previous iterations. To evaluating the gradient magnitude for each sample, an effective scheme is applied to avoid the explicit calculation of the gradient of each sample. This scheme makes it possible that the computation for multiple samples can still be efficiently vectorized. Moreover, the computation cost is only $O(m + l)$ for a $m \times l$ parameter matrix. Therefore, for Active Sampler, the computation overhead introduced in each iteration is light-weight, considering its significant contribution in reducing the number of total iterations.

The contributions of our work are summarized as follows:

- We propose a general SGD accelerator, called Active Sampler, where more informative training data is sampled more frequently for model training. We formalize

Table 1: Examples of ERM Applications

Algorithm	$f_{\mathbf{w}}(\mathbf{x})$	$L(f_{\mathbf{w}}(\mathbf{x}), y)$	$\rho_f(\mathbf{w})$
Linear Regression	$\mathbf{w}^T \mathbf{x}$	$(y - f_{\mathbf{w}}(\mathbf{x}))^2$	0 or $\lambda \ \mathbf{w}\ _2^2$
Hinge-loss SVM	$\mathbf{w}^T \mathbf{x}$	$\max(0, 1 - f_{\mathbf{w}}(\mathbf{x}) * y)$	0 or $\lambda \ \mathbf{w}\ _2^2$
Logistic Regression	$\mathbf{w}^T \mathbf{x}$	$\log(1 + \exp(-f_{\mathbf{w}}(\mathbf{x}) * y))$	0 or $\lambda \ \mathbf{w}\ _2^2$
Feature Selection	$\mathbf{w}^T \mathbf{x}$	$\log(1 + \exp(-f_{\mathbf{w}}(\mathbf{x}) * y))$	$\lambda \ \mathbf{w}\ _1$
Neural Network	complex	$\log(1 + \exp(-f_{\mathbf{w}}(\mathbf{x}) * y))$	0 or $\lambda \ \mathbf{w}\ _2^2$
PCA	$\mathbf{w}\mathbf{w}^T \mathbf{x} - \mathbf{x}$	$\ f_{\mathbf{w}}(\mathbf{x})\ _2^2$	0

the problem as SGD optimization for ERM with weighted sampling, and show that the Active Sampler has the largest information gain and the smallest variance among all weighted sampling solutions.

- We develop a light-weight and fully vectorized algorithm for Active Sampler, making the computation cost of Active Sampler in each iteration comparable to the naive mini-batch SGD.
- We implement the Active Sampler framework and evaluate its performance on three popular machine learning algorithms: SVM, feature selection and deep neural network. Active Sampler reduces the number of iterations to reach a certain accuracy by half, while only consuming 10%-20% additional computation cost in each iteration. In short, Active Sampler speeds up the training procedure by more than 1.6x.

The remainder of this paper is organized as follows. We first introduce the background in Section 2. Then we propose the Active Sampler, show its effectiveness in theory and discuss its practical implementation issues in Section 3. The experimental results are discussed in Section 4. Finally, we review about the related works in Section 5 and conclude at Section 6.

2. PRELIMINARIES

In this section, we first introduce the Empirical Risk Minimization (ERM) framework, and show its connection to data analytics models. Then we describe the stochastic gradient descent algorithm, a general solver for the ERM problem, which we aim to improve upon in this work.

2.1 Empirical Risk Minimization

Empirical Risk Minimization (ERM) is a principle in the statistical learning theory which forms the basis for defining a family of analytics models. From the view of ERM, the central idea in machine learning is to learn a model and use it to approximate the data. The difference between the approximation and the real data is then measured by a loss function, which should be minimized by tuning the parameters of the model. For the sake of simplicity, in this work we formalize ERM from the supervised learning perspective, where each training instance is a pair $\langle \mathbf{x}, y \rangle$ consisting of content \mathbf{x} and label y . For unsupervised problems, label y is a null term Φ , and data can be represented as $\langle \mathbf{x}, \Phi \rangle$.

Definition 1 (Loss Function) *Given a data instance represented as $\langle \mathbf{x}, y \rangle$, and a model hypothesis $f_{\mathbf{w}}$ (i.e. a model f with parameter \mathbf{w}), the loss function $L(f_{\mathbf{w}}(\mathbf{x}), y)$ is a disagreement measure function between the model approximation (i.e. prediction) $f_{\mathbf{w}}(\mathbf{x})$ and the actual label y .*

Here we shall give several examples on the loss measures that are commonly used. For hard classification problems,

the loss measure function can be written as: $L(f_{\mathbf{w}}(\mathbf{x}), y) = I[f_{\mathbf{w}}(\mathbf{x}) \neq y]$. For linear regression, the loss measure function can be written as: $L(f_{\mathbf{w}}(\mathbf{x}), y) = \|f_{\mathbf{w}}(\mathbf{x}) - y\|_2^2$. For general linear models where $f_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^T \mathbf{x}$ and $y = \pm 1$, if the loss measure is $L(f_{\mathbf{w}}(\mathbf{x}), y) = \max(0, 1 - f_{\mathbf{w}}(\mathbf{x}) * y)$, then the model is a hinge-loss SVM; if the loss measure is the log logistic function $L(f_{\mathbf{w}}(\mathbf{x}), y) = \log(1 + \exp(-f_{\mathbf{w}}(\mathbf{x}) * y))$, then the model is a logistic regression. For PCA, by using \mathbf{w} to denote the low rank projection matrix, the problem can be formulated as $f_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}\mathbf{w}^T \mathbf{x} - \mathbf{x}$, and $L(f_{\mathbf{w}}(\mathbf{x}), y) = \|f_{\mathbf{w}}(\mathbf{x})\|_2^2$.

After defining the measure of disagreement between model output and target label, the ultimate goal of learning is naturally to minimize the total disagreement by tuning the model parameters. This is called *Risk Minimization* [26], which is defined as follows:

Definition 2 (Risk Minimization) Let $P_{\langle \mathbf{x}, y \rangle}$ to be the distribution of data, the risk associated with model hypothesis $f_{\mathbf{w}}$ is defined as the expectation of the loss for the potential data distribution:

$$R(f_{\mathbf{w}}) = E[L(f_{\mathbf{w}}(\mathbf{x}), y)] = \int L(f_{\mathbf{w}}(\mathbf{x}), y) dP_{\langle \mathbf{x}, y \rangle} \quad (1)$$

The goal of learning algorithms is to find the parameter \mathbf{w} that minimizes the risk:

$$\underset{\mathbf{w}}{\operatorname{argmin}} R(f_{\mathbf{w}}) \quad (2)$$

However, in general, $R(f_{\mathbf{w}})$ cannot be directly minimized since the exact latent data distribution $P_{\langle \mathbf{x}, y \rangle}$ is unknown. Instead, the common way is to use the distribution of training data to approximate $P_{\langle \mathbf{x}, y \rangle}$. Therefore, the *Empirical Risk* [26] is used as the optimization target.

Definition 3 (Empirical Risk) The empirical risk is defined as the average of loss on the training set with n instances.

$$R_{\text{emp}}(f_{\mathbf{w}}) = \frac{1}{n} \sum_i L(f_{\mathbf{w}}(\mathbf{x}_i), y_i) \quad (3)$$

To simplify the notation, we use $L(\mathbf{w})$ to denote $R_{\text{emp}}(f_{\mathbf{w}})$.

According to the VC-dimension theory [26], the difference between real risk and empirical risk may be large when the model hypothesis $f_{\mathbf{w}}$ is too complex while the size of training data n is not large enough. This phenomenon is called over-fitting. To prevent over-fitting, the empirical risk is often regularized to penalize the complexity of model $f_{\mathbf{w}}$:

Definition 4 (ERM with Regularization) The empirical risk with regularization is defined as the average of loss function on the training set, plus a penalty regularization term $\rho_f(\mathbf{w})$ based on the complexity of the model $f_{\mathbf{w}}$.

$$R_{\text{reg-emp}}(f_{\mathbf{w}}) = L(\mathbf{w}) + \rho_f(\mathbf{w}) \quad (4)$$

In ERM with regularization, the goal of learning algorithm is to minimize the empirical risk with regularization, i.e.,

$$\underset{\mathbf{w}}{\operatorname{argmin}} R_{\text{reg-emp}}(f_{\mathbf{w}}) \quad (5)$$

For most applications, we use the l_2 -norm of parameter $\lambda \|\mathbf{w}\|_2^2$ as the regularization function $\rho_f(\mathbf{w})$. This is actually a Gaussian prior over the parameter distribution from

Table 2: Common Notations

Notation	Meaning
$\langle \mathbf{x}, y \rangle$	training instance
\mathbf{w}	model parameters
$f_{\mathbf{w}}(\mathbf{x})$	model prediction for data \mathbf{x}
$L(f_{\mathbf{w}}(\mathbf{x}), y)$	loss on instance $\langle \mathbf{x}, y \rangle$
$L(\mathbf{w})$	empirical risk
$\rho_f(\mathbf{w})$	regularization term
$\nabla_{\mathbf{w}}$	gradient operator
$\nabla_{\mathbf{w}} L(\mathbf{w})$	batch gradient
$g_i(\mathbf{w})$	stochastic gradient
p_i	sampling probability for $\langle \mathbf{x}_i, y_i \rangle$
$\text{Var}(g_i(\mathbf{w}))$	scalar variance of $g_i(\mathbf{w})$

the Bayesian view. For feature selection methods such as Lasso [16], the l_1 -norm regularization $\lambda \|\mathbf{w}\|_1$ is used to select those sparse features.

We list some examples of analytics models from ERM family in Table 1, and show their connections.

2.2 Stochastic Gradient Descent

To optimize the ERM problem described in Equation 5, batch gradient descent method is used to iteratively alter the parameter towards the fastest direction to minimize the objective function. By defining the step size using the learning rate η , batch gradient descent method uses the following updating rule to optimize the parameter:

$$\begin{aligned} \mathbf{w}_{\text{new}} &= \mathbf{w} - \eta \nabla_{\mathbf{w}} L(\mathbf{w}) + \rho_f(\mathbf{w}) \\ &= \mathbf{w} - \eta \nabla_{\mathbf{w}} \rho_f(\mathbf{w}) - \frac{\eta}{n} \sum_i \nabla_{\mathbf{w}} L(f_{\mathbf{w}}(\mathbf{x}_i), y_i) \end{aligned} \quad (6)$$

As can be observed from Equation 6, we need to evaluate the gradients $\nabla_{\mathbf{w}} L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)$ for all training instances at each step, making the computation cost of $\nabla_{\mathbf{w}} L(\mathbf{w})$ extremely expensive. To avoid this cost, stochastic gradient methods use an inexact gradient which is estimated from random samples.

Definition 5 (Stochastic Gradient Descent) In stochastic gradient descent, the true gradient $\nabla_{\mathbf{w}} L(\mathbf{w})$ is approximated by a stochastic gradient $g_i(\mathbf{w})$.

$$\mathbf{w}_{\text{new}} = \mathbf{w} - \eta \nabla_{\mathbf{w}} \rho_f(\mathbf{w}) - \eta g_i(\mathbf{w}) \quad (7)$$

Taking $g_i(\mathbf{w})$ as a random variable, the expectation of $g_i(\mathbf{w})$ should equal to the gradient of $L(\mathbf{w})$, i.e.

$$E_i[g_i(\mathbf{w})] = \nabla_{\mathbf{w}} L(\mathbf{w}) \quad (8)$$

In the standard SGD algorithm, $g_i(\mathbf{w})$ is obtained by simply evaluating the gradient at a random single instance i :

$$g_i(\mathbf{w}) = \nabla_{\mathbf{w}} L(f_{\mathbf{w}}(\mathbf{x}_i), y_i) \quad (9)$$

where i is randomly drawn from $\{1, \dots, n\}$, and the sampling probability p_i for each instance i is $1/n$. The scalar variance of stochastic gradient is denoted as $\text{Var}(g_i(\mathbf{w}))$, defined by $E_i[\|g_i(\mathbf{w}) - \nabla_{\mathbf{w}} L(\mathbf{w})\|_2^2]$, which is a scalar instead of the covariance matrix.

Table 2 lists most of the important notations used throughout this paper. Unless otherwise specified, variance used in the paper refers to the scalar variance. Intuitively, the requirement $E_i[g_i(\mathbf{w})] = \nabla_{\mathbf{w}} L(\mathbf{w})$ is to guarantee that the

SGD algorithms will converge at the optimal point [20], as the expectation of update in SGD will be a zero vector at the point where batch gradient descent algorithm converges. Obviously, the standard SGD algorithm satisfies this requirement.

$$\begin{aligned}
E_i[g_i(\mathbf{w})] &= \sum_i p_i g_i(\mathbf{w}) \\
&= \sum_i \frac{1}{n} \nabla_{\mathbf{w}} L(f_{\mathbf{w}}(\mathbf{x}_i), y_i) \\
&= \nabla_{\mathbf{w}} \frac{1}{n} \sum_i L(f_{\mathbf{w}}(\mathbf{x}_i), y_i) \\
&= \nabla_{\mathbf{w}} L(\mathbf{w})
\end{aligned} \tag{10}$$

Although the training cost per iteration for SGD is extremely light-weight compared to the batch gradient algorithm, the main drawback of SGD is that $g_i(\mathbf{w})$ is not the exact $\nabla_{\mathbf{w}} L(\mathbf{w})$. Therefore, the direction of the stochastic gradient $g_i(\mathbf{w})$ differs from the optimal direction $\nabla_{\mathbf{w}} L(\mathbf{w})$. This phenomenon causes SGD algorithm to be less efficient and take more iterations to converge or reach a certain accuracy. It has been shown by [27, 11, 23] that simply reducing the variance of stochastic gradient will increase the convergence rate of SGD algorithms. However, most variance reduction techniques require additional computation cost compared with the standard SGD. In essence, there is a trade-off between the number of iterations required to reach a certain accuracy and the computational cost per iteration. Therefore, the main goal of optimizing SGD algorithm is to reduce the variance of $g_i(\mathbf{w})$, while keeping the computation of $g_i(\mathbf{w})$ light-weight.

3. ACTIVE SAMPLER

3.1 Overview

In this paper, we revisit the SGD algorithm from a brand new angle – the information gain of the model at each iteration. We can regard the SGD algorithm as an active learning procedure which sequentially takes samples from the dataset and refines its model. Naturally, training the model using samples with more information would facilitate faster improvement of the model. We term our weighted sampling strategy as *Active Sampler*. The intuition here is that a larger number of training samples are not helpful for refining the model (or at least not helpful at a certain training stage), including data that are too evident to predict, data that are noisy, and data that have just been visited. By simply skipping these samples, we can save a significant amount of training time. In contrast, the samples that are close to the border of class may be very helpful to refine the model (or even define the model in some cases such as SVM). This idea is very similar to active learning but with a major difference – the objective of active learning is to reduce the number of training samples, while the objective of active sampling is to reduce the number of training iterations.

To exploit information gain as a means to speed up SGD training, three issues have to be addressed: (1) define what is the information gain for model training from each training sample; (2) adapt the Active Sampler into the SGD framework and study how information gain can help speeding up SGD; (3) design a light-weight implementation that can be

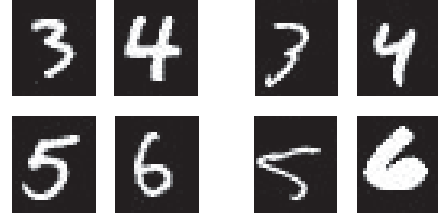


Figure 2: Uncertainty

applied to real systems. We shall address these three issues in the following subsections.

3.2 Information Gain

In this subsection, we define the information gain directly following the basic intuition from the training of typical soft margin classifiers. Later, we will fit this initial idea into a formal SGD framework and provide a rigorous theoretical analysis that illustrates how this strategy can benefit the performance of SGD for all sorts of ERM problems in the next subsection.

In soft margin classifiers, instead of giving a single label as the prediction, the classifier outputs a probabilistic distribution over latent labels. Recently proposed classification algorithms (e.g. Lasso, Neural Network and Soft SVM etc.) are typically trained as soft margin classifiers, since the continuous optimization is much efficient than the discrete optimization, which is a NP-Hard problem.

Definition 6 (Soft Margin Classifier) *Given a predictor $f_{\mathbf{w}}$, the soft probabilistic classifier is defined by using log logistic function as the loss function, i.e.*

$$L(f_{\mathbf{w}}(\mathbf{x}), y) = \log(1 + \exp(-f_{\mathbf{w}}(\mathbf{x}) * y)) \tag{11}$$

Obviously, all algorithms in Table 1 using the logistic loss function are soft margin classifiers. The rationale is that the logistic function is used to transfer the prediction $f_{\mathbf{w}} \in R$ to a classification probability, i.e.

$$Pr(y|f_{\mathbf{w}}(\mathbf{x})) = 1/(1 + \exp(-f_{\mathbf{w}}(\mathbf{x}) * y)) \tag{12}$$

*Then the log-likelihood $\log Pr(y|f_{\mathbf{w}}(\mathbf{x}))$ is maximized, i.e. the loss is $\log(1 + \exp(-f_{\mathbf{w}}(\mathbf{x}) * y))$.*

Now, let us analyze the possible factors that may affect the information gain of a model from each training sample. First, from the view of active learning, the information gain by revealing a label which can be easily predicted is very limited. This is also true for the SGD algorithm – visiting a sample that the model can always classify correctly is not helpful, as the loss cannot be further reduced if the loss is already close to 0. A model can only learn from samples which are uncertain in prediction. Figure 2 shows an example to illustrate why uncertainty helps the model to improve: the images on the left are very easy to predict, and therefore $Pr(y|f_{\mathbf{w}}(\mathbf{x}))$ is almost 100%. As a result, the loss $L(f_{\mathbf{w}}(\mathbf{x}), y)$ for each of those images is almost zero and has little room for further optimization. In contrast, the images on the right are much harder to predict. By selecting them as samples for optimization purpose, the loss $L(f_{\mathbf{w}}(\mathbf{x}), y)$ on those samples could be significantly reduced and the average performance of model is hence improved.



Figure 3: Significance

In information theory [26], the information gain by revealing a random variable is usually defined as the entropy of that random variable:

Definition 7 (Uncertainty) *The uncertainty of a training instance \mathbf{x}_i for a model $f_{\mathbf{w}}$ is defined as the entropy of the $f_{\mathbf{w}}(\mathbf{x}_i)$, i.e.*

$$U(\mathbf{w}, \mathbf{x}_i) = - \sum_y Pr(y|f_{\mathbf{w}}(\mathbf{x}_i)) \log Pr(y|f_{\mathbf{w}}(\mathbf{x}_i)) \quad (13)$$

Second, not all the training instances contribute equally to the model performance. For example, though the model may always be uncertain about the label for noisy data, this does not mean that noisy data are more helpful to improve the model performance. This is because the uncertainty measure only evaluates the information inside the label, but not its contribution to the model. Therefore, we introduce another measure called *significance* to evaluate the efficiency of information transfer from the data to the model. Intuitively, the output of a noisy instance is not sensitive to the change of the parameter. When the output of a data instance is sensitive to the change of parameter, its loss will be significantly reduced even with tiny changes of the parameter, which provides a clear instruction on how to reduce the loss by tuning the parameter. The right hand side of Figure 3 shows the images that are noisy and with less significance.

Definition 8 (Significance) *The significance of a training instance \mathbf{x}_i for a model $f_{\mathbf{w}}$ is defined as its sensitivity to parameter change:*

$$S(\mathbf{w}, \mathbf{x}_i) = \|\nabla_{\mathbf{w}} f_{\mathbf{w}}(\mathbf{x}_i)\|_2 \quad (14)$$

$\|\nabla_{\mathbf{w}} f_{\mathbf{w}}(\mathbf{x}_i)\|_2$ is the maximal change of $f_{\mathbf{w}}(\mathbf{x}_i)$ when the parameter \mathbf{w} changes an unit distance.

Here, we give a comparison between uncertainty and significance: uncertainty measures the expectation of accuracy on the current model, while significance measures the potential improvement of accuracy by tuning the model. Therefore, instances that are easy to classify usually have low uncertainty but high significance; noisy instances usually have high uncertainty but low significance, while valuable border instances that have not been well learned usually have both high uncertainty and high significance.

Third, the information gain in one iteration may overlap with the information obtained in earlier training steps. For example, visiting the training instance that has just been trained usually does not provide extra information than what has been derived in the previous visit. However, for a completely new instance that hasn't been trained, there may be no overlap between the information gain and information

in previous steps. Therefore, we use the visiting interval to measure the effect of information overlap:

Definition 9 (Interval) *The visiting interval $I(\mathbf{w}, \mathbf{x}_i)$ of a training instance \mathbf{x}_i for a model $f_{\mathbf{w}}$ is defined as the number of iterations since the last time \mathbf{x}_i was used in training. A larger interval provides less information overlap and more pure information gain.*

Combining all the three factors together, we define the information gain of model $f_{\mathbf{w}}$ from training instance \mathbf{x}_i as $IG(\mathbf{w}, \mathbf{x}_i)$:

$$IG(\mathbf{w}, \mathbf{x}_i) = U(\mathbf{w}, \mathbf{x}_i) * S(\mathbf{w}, \mathbf{x}_i) * I(\mathbf{w}, \mathbf{x}_i) \quad (15)$$

The objective of our Active Sampler is to choose the training instance \mathbf{x}_i with the largest $IG(\mathbf{w}, \mathbf{x}_i)$.

Theorem 1 (Information Gain Maximization) *By choosing the largest $IG(\mathbf{w}, \mathbf{x}_i)$ in each iteration, the sampling frequency p_i of each training instance \mathbf{x}_i should be proportional to its expectation of the gradient magnitude, i.e.*

$$p_i = \frac{E_y[\|\nabla_{\mathbf{w}} L(f_{\mathbf{w}}(\mathbf{x}_i), y)\|_2]}{\sum_i E_y[\|\nabla_{\mathbf{w}} L(f_{\mathbf{w}}(\mathbf{x}_i), y)\|_2]} \quad (16)$$

PROOF SKETCH: At each iteration, $IG(\mathbf{w}, \mathbf{x}_i)$ for each instance will increase $U(\mathbf{w}, \mathbf{x}_i) * S(\mathbf{w}, \mathbf{x}_i)$, and the instance with the largest $IG(\mathbf{w}, \mathbf{x}_i)$ will be selected as sample. After an instance is sampled, its $IG(\mathbf{w}, \mathbf{x}_i)$ will be set to zero as its $I(\mathbf{w}, \mathbf{x}_i)$ becomes zero. Therefore, as the number of iterations grows, the sampling frequency for each instance should be proportional to its $U(\mathbf{w}, \mathbf{x}_i) * S(\mathbf{w}, \mathbf{x}_i)$, considering that the largest $IG(\mathbf{w}, \mathbf{x}_i)$ selected in each iteration should have a similar value. Meanwhile,

$$\begin{aligned} & E_y[\|\nabla_{\mathbf{w}} L(f_{\mathbf{w}}(\mathbf{x}_i), y)\|_2] \\ = & \sum_y Pr(y|f_{\mathbf{w}}(\mathbf{x}_i)) \|\nabla_{\mathbf{w}} L(f_{\mathbf{w}}(\mathbf{x}_i), y)\|_2 \\ = & \sum_y Pr(y|f_{\mathbf{w}}(\mathbf{x}_i)) \frac{\partial}{\partial f_{\mathbf{w}}(\mathbf{x}_i)} L(f_{\mathbf{w}}(\mathbf{x}_i), y) \|\nabla_{\mathbf{w}} f_{\mathbf{w}}(\mathbf{x}_i)\|_2 \\ = & \sum_y (Pr(y|f_{\mathbf{w}}(\mathbf{x}_i)) * \log Pr(y|f_{\mathbf{w}}(\mathbf{x}_i))) \|\nabla_{\mathbf{w}} f_{\mathbf{w}}(\mathbf{x}_i)\|_2 \\ = & U(\mathbf{w}, \mathbf{x}_i) * S(\mathbf{w}, \mathbf{x}_i) \end{aligned} \quad (17)$$

□

3.3 Weighted SGD Algorithm and Analysis

While the above intuition suggests that different training instances should be sampled at different frequencies, directly changing the sampling frequency will result in a bias in the target of an optimization.

$$\begin{aligned} E_i[g_i(\mathbf{w})] &= \sum_i p_i g_i(\mathbf{w}) \\ &= \nabla_{\mathbf{w}} \sum_i p_i L(f_{\mathbf{w}}(\mathbf{x}_i), y_i) \end{aligned} \quad (18)$$

The loss function to be minimized is $\sum_i p_i L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)$ instead of $\sum_i \frac{1}{n} L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)$. The weight for each training instance is unequal, which may affect the accuracy of the model. For this reason, using $\nabla_{\mathbf{w}} L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)$ directly as $g_i(\mathbf{w})$ is inappropriate. Instead, we should guarantee that $E_i[g_i(\mathbf{w})]$ is $\nabla_{\mathbf{w}} L(\mathbf{w})$.

Theorem 2 (Weighted SGD) *Given any sampling distribution $\{p_1, \dots, p_i, \dots, p_n\}$, to get a SGD algorithm that optimizes $L(\mathbf{w})$, $g_i(\mathbf{w})$ should be re-weighted to $\frac{\nabla_{\mathbf{w}} L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)}{n * p_i}$.*

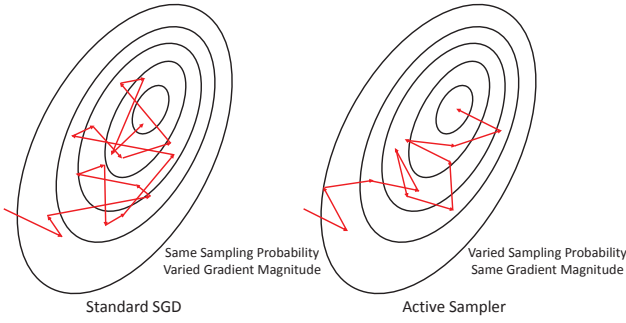


Figure 4: Comparison with standard SGD

PROOF. To get a SGD algorithm that optimizes $L(\mathbf{w})$, we need $E_i[g_i(\mathbf{w})]$ to be $\nabla_{\mathbf{w}}L(\mathbf{w})$. By scaling $g_i(\mathbf{w})$ w_i times and solve $E_i[g_i(\mathbf{w})] = \nabla_{\mathbf{w}}L(\mathbf{w})$, we have:

$$\begin{aligned}
 E_i[g_i(\mathbf{w})] &= \nabla_{\mathbf{w}}L(\mathbf{w}) \\
 \Rightarrow E_i[w_i \nabla_{\mathbf{w}}L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)] &= \nabla_{\mathbf{w}}L(\mathbf{w}) \\
 \Rightarrow \sum_i p_i w_i \nabla_{\mathbf{w}}L(f_{\mathbf{w}}(\mathbf{x}_i), y_i) &= \sum_i \frac{1}{n} \nabla_{\mathbf{w}}L(f_{\mathbf{w}}(\mathbf{x}_i), y_i) \\
 \Rightarrow p_i w_i &= \frac{1}{n} \quad (19)
 \end{aligned}$$

Therefore, $w_i = 1/(n * p_i)$. \square

Next, we show that setting p_i proportional to the gradient magnitude will minimize the variance in stochastic gradient $g_i(\mathbf{w})$ in all weighted sampling solutions that $E_i[g_i(\mathbf{w})] = \nabla_{\mathbf{w}}L(\mathbf{w})$.

Theorem 3 (Optimal Weighted SGD) Let p_i denote the sampling probability of training instance $\langle \mathbf{x}_i, y_i \rangle$ in a weighted SGD algorithm where $E_i[g_i(\mathbf{w})] = \nabla_{\mathbf{w}}L(\mathbf{w})$. $g_i(\mathbf{w})$ becomes a stochastic gradient with value $\frac{\nabla_{\mathbf{w}}L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)}{n * p_i}$ with sampling probability p_i . To get the weighted SGD algorithm with the smallest variance of stochastic gradient (i.e. $\text{Var}(g_i(\mathbf{w}))$), for each instance, p_i should be proportional to its magnitude of stochastic gradient $\|\nabla_{\mathbf{w}}L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)\|_2$.

$$p_i = \frac{\|\nabla_{\mathbf{w}}L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)\|_2}{\sum_i \|\nabla_{\mathbf{w}}L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)\|_2} \quad (20)$$

PROOF.

$$\begin{aligned}
 &\text{Var}(g_i(\mathbf{w})) \\
 = & E_i[\|g_i(\mathbf{w})\|_2^2] - \|E_i[g_i(\mathbf{w})]\|_2^2 \\
 = & E_i\left[\frac{\|\nabla_{\mathbf{w}}L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)\|_2^2}{(np_i)^2}\right] - \|E_i\left[\frac{\nabla_{\mathbf{w}}L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)}{np_i}\right]\|_2^2 \\
 = & \sum_i p_i \frac{\|\nabla_{\mathbf{w}}L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)\|_2^2}{(np_i)^2} - \left\|\sum_i p_i \frac{\nabla_{\mathbf{w}}L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)}{np_i}\right\|_2^2 \\
 = & \sum_i \frac{\|\nabla_{\mathbf{w}}L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)\|_2^2}{n^2 p_i} - \|\nabla_{\mathbf{w}}L(\mathbf{w})\|_2^2 \quad (21)
 \end{aligned}$$

To minimize $\text{Var}(g_i(\mathbf{w}))$ by tuning p_i , subjecting to $\sum_i p_i = 1$, according to Lagrange multiplier method, we have:

$$\begin{cases} \frac{\partial}{\partial p_i}(\text{Var}(g_i(\mathbf{w})) + \lambda(\sum_i p_i - 1)) = 0 \\ \frac{\partial}{\partial \lambda}(\text{Var}(g_i(\mathbf{w})) + \lambda(\sum_i p_i - 1)) = 0 \end{cases} \quad (22)$$

by solving Equation 22, we have:

$$\lambda - \frac{\|\nabla_{\mathbf{w}}L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)\|_2^2}{(np_i)^2} = 0 \quad (23)$$

Algorithm 1: Optimal Weighted SGD

Input: Initial \mathbf{w}_0 , T
Output: Final \mathbf{w}_T

```

1 for  $t = 1, \dots, T$  do
2   foreach  $i = 1, \dots, n$  do
3      $\text{Grad}[i] = \|\nabla_{\mathbf{w}}L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)\|_2$ ;
4    $\text{SumGrad} = \sum_i \text{Grad}[i]$ ;
5   foreach  $i = 1, \dots, n$  do
6      $p_i = \text{Grad}[i] / \text{SumGrad}$ ;
7   sample  $i$  from  $\{1, \dots, n\}$  based on distribution  $\{p_1, \dots, p_n\}$ ;
8    $g_i(\mathbf{w}) = \nabla_{\mathbf{w}}L(f_{\mathbf{w}}(\mathbf{x}_i), y_i) / np_i$ ;
9    $\mathbf{w}_t = \mathbf{w}_{t-1} - \eta \nabla_{\mathbf{w}}\rho_f(\mathbf{w}) - \eta g_i(\mathbf{w})$ ;

```

Therefore, $\frac{\|\nabla_{\mathbf{w}}L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)\|_2^2}{(np_i)^2}$ is a constant value for all training instances, which means $\|\nabla_{\mathbf{w}}L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)\|_2$ is proportional to p_i . Considering that $\sum_i p_i = 1$, we have:

$$p_i = \frac{\|\nabla_{\mathbf{w}}L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)\|_2}{\sum_i \|\nabla_{\mathbf{w}}L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)\|_2} \quad (24)$$

\square

Theorem 3 can be viewed as a refined version of Theorem 1 when label y is observed. Theorem 3 gives a rigorous explanation about our previous intuition from the variance reduction view of SGD optimization. Note that this result makes no assumptions about the soft margin classification and applies to all sorts of ERM problems. Algorithm 1 describes the optimal weighted SGD algorithm, where the computation cost in each iteration will be optimized in the next subsection.

Another insight we observe from Theorem 3 is that in order to minimize the variance of SGD by using weighted sampling, $g_i(\mathbf{w})$ should have the exact same magnitude for all instances.

$$\begin{aligned}
 \|g_i(\mathbf{w})\|_2 &= \left\|\frac{\nabla_{\mathbf{w}}L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)}{np_i}\right\|_2 \\
 &= \frac{\|\nabla_{\mathbf{w}}L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)\|_2}{n * \frac{\|\nabla_{\mathbf{w}}L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)\|_2}{\sum_i \|\nabla_{\mathbf{w}}L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)\|_2}} \\
 &= \frac{\sum_i \|\nabla_{\mathbf{w}}L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)\|_2}{n} \quad (25)
 \end{aligned}$$

This suggests that only the **direction** of the stochastic gradient is determined by the training instance. Further, the step size (i.e. the magnitude of $\eta g_i(\mathbf{w})$) is of no consequence to the training instance and is decided by the global learning rate. As a result, the change of parameter in Active Sampler is much more steady than the standard SGD. This property agrees with the original purpose of gradient descent methods, as the gradient only indicates the fastest direction to minimize the objective loss function, without any indication on the step size. Figure 4 shows the comparison between standard SGD and Active Sampler. For standard SGD, all training instances have the same sampling probability, while their gradient magnitudes vary. For Active Sampler, all training instances have the same gradient magnitude, while their sampling probabilities vary. Both methods have the same expectation of gradient, however, Active Sampler

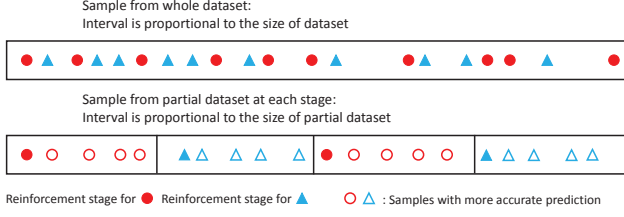


Figure 5: History Reinforcement

has a smaller variance, resulting in a faster and more stable convergence process.

3.4 Practical Implementation Issues

As discussed in Section 2, the main goal of optimizing the SGD algorithm is to reduce the variance of $g_i(\mathbf{w})$, while keeping the computation cost per iteration light-weight. We have already shown how to minimize the variance of $g_i(\mathbf{w})$ by using the Active Sampler. In this subsection, we will discuss some practical issues in implementing Active Sampler onto real systems, which may significantly affect the computation time in each iteration.

3.4.1 Sampling based on History

The probabilistic distribution described in Theorem 3 can indeed minimize the variance of the stochastic gradient. However, use of the exact distribution, which requires all n gradients to be evaluated at each step, is obviously not practical. Instead, we can predict the gradient magnitude for each training instances using historical data. A straightforward approach to the problem is to remember the magnitude of the latest gradient of each instance and use it as an approximation. Considering that the actual gradient may change and the historical magnitude is only an approximation, a smoothing term is required. For example, if one instance contributes a zero gradient at any iteration of the model training when it is sampled, that sample will never be visited afterward if there is no smoothing, notwithstanding this instance may become valuable for refinement of the model at later stages.

Definition 10 (History Approximation) Let t_i be the latest step where training instance i is visited and let \mathbf{w}_{t_i} denote the parameter value at step t_i . The sampling probability for each sample i in a practical Active Sampler with smoothing is defined as:

$$p_i = (1 - \beta) \frac{\|\nabla_{\mathbf{w}_{t_i}} L(f_{\mathbf{w}_{t_i}}(\mathbf{x}_i), y_i)\|_2}{\sum_i \|\nabla_{\mathbf{w}_{t_i}} L(f_{\mathbf{w}_{t_i}}(\mathbf{x}_i), y_i)\|_2} + \frac{\beta}{n} \quad (26)$$

The scheme ensures that every training instance has at least β times the average sampling probability (i.e. $1/n$) being sampled. Algorithm 2 describes the Active Sampler using the history approximation. In each iteration, only its gradient magnitude needs to be updated.

We note that by using the history length to denote the number of iterations from the last time an instance is sampled, its history approximation becomes less accurate with the increase of the history length. Meanwhile, the expectation of history length for one instance is $1/p_i$, and the average of p_i is $1/n$. Consequently, the history approximation will become less accurate when data size becomes larger. To

Algorithm 2: ASSGD (Active Sampler SGD)

Input: Initial \mathbf{w}_0 , T , β , $Grad[]$,
 $SumGrad = \sum_i Grad[i]$
Output: Final \mathbf{w}_T

```

1 for  $t = 1, \dots, T$  do
2   sample  $i$  from  $\{1, \dots, n\}$  based on distribution
    $\{p_1, \dots, p_n\}$  where
    $p_i = \beta/n + (1 - \beta)Grad[i]/SumGrad$ ;
3    $g_i(\mathbf{w}) = \nabla_{\mathbf{w}} L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)/np_i$ ;
4    $\mathbf{w}_t = \mathbf{w}_{t-1} - \eta \nabla_{\mathbf{w}} \rho_f(\mathbf{w}) - \eta g_i(\mathbf{w})$ ;
5    $SumGrad = SumGrad - Grad[i]$ ;
6    $Grad[i] = \|\nabla_{\mathbf{w}} L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)\|_2$ ;
7    $SumGrad = SumGrad + Grad[i]$ ;

```

address this issue, we propose *History Reinforcement*, whose key idea is illustrated in Figure 5. History Reinforcement algorithm trains the model using a set of stages, each of which contains a large amount of SGD iterations. Within a stage, it first samples a subset of training data which consists of m instances, and then uses them as the training set in its SGD iterations. During the training of each stage, the sampling probability for the instances is n/m times larger than training all the instances together. Therefore, the approximation will be much more accurate except the first time in a stage when one instance is sampled. The only drawback of History Reinforcement is that it may lead to a bias in the training of a stage, as only partial data are trained. However, [15] presents an effective scheme in an analogous context to reduce this bias by adding a regularizer to limit the change of parameter in one stage. Below, we formally define the concept of **History Reinforcement**.

Definition 11 (History Reinforcement) *History Reinforcement algorithm consists multiple stages. in each stage t , it draws a subset I_t of training instances, which contains m random instances from the whole dataset, and trains the model \mathbf{w}_t using g SGD iterations. The loss function used in each stage is:*

$$L(\mathbf{w}_t) = \sum_{i \in I_t} \frac{L(f_{\mathbf{w}_t}(\mathbf{x}_i), y_i)}{m} + \frac{\gamma_t}{2} \|\mathbf{w}_{t-1} - \mathbf{w}_t\|_2^2 \quad (27)$$

where γ_t is a parameter in [15] calculated based on m , t and $Var(g_i(\mathbf{w}))$.

The correctness and effectiveness of this batch training is given in Theorem 1 of [15] (by considering a stage as a batch step). The average number of visits for one instance in a stage is g/m . Therefore, $1 - m/g$ proportion of the iterations in a training stage will benefit from a more accurate approximation. In essence, there is a trade-off between the bias involved by using partial data and the accuracy gain in gradient approximation. Intuitively, for larger datasets, the bias becomes less significant while the accuracy gain by using History Reinforcement becomes more valuable. On the contrary, the approximation of gradient in small datasets is fairly accurate, and therefore, directly sampling from the whole dataset is advantageous.

3.4.2 Efficient Vectorized Computation

To reduce the variance of the stochastic gradient, a widely adopted solution is to employ mini-batch training, which av-

Algorithm 3: ASSGD with History Reinforcement

Input: Initial \mathbf{w}_0 , T , m , g
Output: Final \mathbf{w}_T

```

1 for  $t = 1, \dots, T$  do
2    $I_t = \Phi$ ;
3   for  $i = 1, \dots, m$  do
4     sample  $t_i$  uniformly from  $\{1, \dots, n\} - I_t$ ;
5      $I_t = I_t \cup \{t_i\}$ ;
6   Compute  $\gamma_t$  based on [15];
7   Train  $\mathbf{w}_t$  using Algorithm 2 for  $g$  iterations, using
    $\mathbf{w}_{t-1}$  as initial  $\mathbf{w}_0$ , using  $I_t$  as the training set, and
   using  $\rho_f(\mathbf{w}) + \frac{\gamma_t}{2} \|\mathbf{w}_{t-1} - \mathbf{w}\|_2^2$  as the regularization
   function;
```

erages the stochastic gradients of multiple training samples. By averaging b training samples, the variance of gradient can be reduced by b times (b is typically between 10 and 1000). Meanwhile, thanks to the effect of vectorized computation and the constant communication cost when the computations are parallelized, the training time per iteration for mini-batch SGD is much smaller than b times the training time per iteration for the standard SGD. Therefore, mini-batch SGD is commonly used in most large-scale optimization problems. Active Sampling is orthogonal to mini-batch SGD, so we could use both improvements simultaneously. However, to integrate them together, we need to compute the average of $g_i(\mathbf{w})$ for b training samples in an efficient vectorized way, as well as to obtain the gradient magnitude for each training instance.

Definition 12 (Mini-batch SGD / Active Sampler)

At each iteration t , mini-batch SGD uniformly draws b samples $I_t = \{t_1, \dots, t_b\}$ from $\{1, \dots, n\}$, and uses the averaged gradient as the stochastic gradient.

$$g_t(\mathbf{w}) = \sum_{i \in I_t} \frac{\nabla_{\mathbf{w}} L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)}{b} \quad (28)$$

At each iteration t , mini-batch Active Sampler repeats the sample selection in Theorem 3 for b times and get b samples $I_t = \{t_1, \dots, t_b\}$, and uses the averaged gradient as the stochastic gradient.

$$g_t(\mathbf{w}) = \sum_{i \in I_t} \frac{\nabla_{\mathbf{w}} L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)}{bnp_i} \quad (29)$$

Similar to mini-batch SGD, the variance of $g_t(\mathbf{w})$ in mini-batch Active Sampler is reduced by b times.

In mini-batch SGD, the main advantage stemmed from vectorized computation is that the actual gradients from all samples do not need to be stored individually and then aggregated. This trick is very time and memory efficient when the size of parameters is huge (e.g. deep neural network, sparse logistic regression). Here we use a multi-layer perceptron [4] (MLP) model to illustrate how the stochastic gradient of mini-batch SGD is computed, and how mini-batch Active Sampler can be computed in a similar light-weight manner. Note that general linear models ($f_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^T \mathbf{x}$) are usually generalized as a multi-class classification problem, and their parameters \mathbf{w} are also a matrix, which is similar to the hidden layer in MLP. Therefore, general linear models can be viewed as a single layer perceptron with

Algorithm 4: Batch Computation for Active Sampler

Input: $H_{b \times l}^{(k)}$, $Z_{b \times m}^{(k+1)}$, $W_{m \times l}^{(k)}$, $\nabla H_{b \times m}^{(k+1)}$
Output: $\nabla H_{b \times l}^{(k)}$, $\nabla W_{m \times l}^{(k)}$, $\|\nabla_{W^{(k)}} L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)\|_2$

```

1 foreach  $i \in \{0, \dots, b-1\}$  do
2   foreach  $p \in \{0, \dots, m-1\}$  do
3      $\nabla Z^{(k+1)}[i][p] = \sigma'(Z^{(k+1)}[i][p]) \nabla H^{(k+1)}[i][p]$ ;
4    $\nabla H_{b \times l}^{(k)} = \nabla Z_{b \times m}^{(k+1)} \times W_{m \times l}^{(k)}$ ;
5    $\nabla W_{m \times l}^{(k)} = \frac{1}{b} (\nabla Z^{(k+1)})_{m \times b}^T \times H_{b \times l}^{(k)}$ ;
6 // line 1-5 : compute stochastic gradient  $O(bml)$ 
7 foreach  $i \in \{0, \dots, b-1\}$  do
8    $SumGZ = 0$ ,  $SumH = 0$ ;
9   foreach  $p \in \{0, \dots, m-1\}$  do
10     $SumGZ = SumGZ + (\nabla Z^{(k+1)}[i][p])^2$ ;
11    foreach  $q \in \{0, \dots, l-1\}$  do
12       $SumH = SumH + (H^{(k)}[i][q])^2$ ;
13     $\|\nabla_{W^{(k)}} L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)\|_2 = \sqrt{SumGZ * SumH}$ 
14 // line 7-13 : compute gradient magnitude  $O(b(m+l))$ 
```

a small difference in the loss function and hence all the optimization techniques discussed below can be applied to these models as well.

Definition 13 (Multi-Layer Perceptron (MLP)) *Multi-Layer Perceptron [4] is a feed forward neural network. It consists of one input layer $H^{(0)}$, h hidden layers ($H^{(k)}$, $k = 1, \dots, h$) and a loss layer to compute the loss $L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)$ based on the prediction $H^{(h)}$ for \mathbf{x}_i . Each hidden layer k is a vector of units, and the calculation is formalized as follows:*

$$Z^{(k+1)} = W^{(k)} H^{(k)} + B^{(k)} \quad (30)$$

$$H^{(k+1)} = \sigma(Z^{(k+1)}) \quad (31)$$

where $\sigma(\cdot)$ is the activation function. The gradient is computed via back-propagation:

$$\frac{\partial L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)}{\partial W^{(k)}} = \frac{\partial L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)}{\partial Z_i^{(k+1)}} \times H_i^{(k)T} \quad (32)$$

$$\frac{\partial L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)}{\partial H^{(k)}} = \frac{\partial L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)}{\partial Z_i^{(k+1)}} \times W^{(k)} \quad (33)$$

$$\frac{\partial L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)}{\partial Z_p^{(k)}} = \sigma'(Z_p^{(k)}) \frac{\partial L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)}{\partial H_p^{(k)}} \quad (34)$$

We now analyze the computation of gradient for one layer k in mini-batch SGD. Using m to denote the number of units in $H^{(k+1)}$, and l to denote the number of units in $H^{(k)}$, the parameter $W^{(k)}$ is an $m \times l$ matrix, and $g_t(W^{(k)})$ is also an $m \times l$ matrix.

$$\begin{aligned}
g_t(W^{(k)}) &= \sum_{i \in I_t} \frac{\nabla_{W^{(k)}} L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)}{b} \\
&= \frac{1}{b} \sum_{i \in I_t} \frac{\partial L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)}{\partial Z_i^{(k+1)}} \times H_i^{(k)T} \quad (35)
\end{aligned}$$

However, directly computing the b gradients $\frac{\partial L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)}{\partial Z_i^{(k+1)}} \times H_i^{(k)T}$ one by one is not efficient, as each gradient is an $m \times l$ matrix. Instead, using $H^{(k)}$ to denote the $b \times l$ lower layer

Table 3: Datasets and Models

Dataset	# Examples	Size	Model	Test Error
MNIST	60K	57MB	kernel SVM	0.6%
URL	2.4M	950MB	Lasso	2.5%
CIFAR10	60K	161MB	DCNN	18%
CIFAR-DA	7.6M	14.8GB	DCNN	11.5%

feature matrix $[H_1^{(k)}, \dots, H_b^{(k)}]^T$, and using $\frac{\partial L(f_{\mathbf{w}}(\mathbf{x}), y)}{\partial Z^{(k+1)}}$ to denote the $b \times m$ higher layer gradient matrix $[\frac{\partial L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)}{\partial Z_1^{(k+1)}}, \dots, \frac{\partial L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)}{\partial Z_b^{(k+1)}}]^T$, we have:

$$g_t(W_{pq}^{(k)}) = \frac{1}{b} \sum_{i \in I_t} \frac{\partial L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)}{\partial Z_{ip}^{(k+1)}} \times H_{iq}^{(k)}$$

$$\Rightarrow g_t(W^{(k)}) = \frac{1}{b} \left(\frac{\partial L(f_{\mathbf{w}}(\mathbf{x}), y)}{\partial Z^{(k+1)}} \right)^T \times H^{(k)} \quad (36)$$

Therefore, it is computed by performing matrix multiplication for an $m \times b$ matrix and a $b \times l$ matrix, which is obviously more efficient than the previous method, which computes multiple vector-vector multiplications. It also reduces the memory cost from $b \times m \times l$ to $m \times l$. The computation for $\frac{\partial L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)}{\partial H^{(k)}}$ is analogous.

In mini-batch Active Sampler, there are two differences compared to mini-batch SGD. First, Active Sampler needs to provide each instance a weight based on $1/np_i$. Second, Active Sampler needs to compute the gradient magnitude for every training instance. For the first problem, the solution is quite straightforward – putting the weight in the loss function before calculating its gradient for the parameters. By scaling the value of loss by $1/np_i$ times, its gradient will change $1/np_i$ times accordingly.

For the calculation of the gradient magnitude, we exploit the following equation to avoid explicitly calculating the gradient of each training instance:

$$\begin{aligned} & \|\nabla_{W^{(k)}} L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)\|_2^2 \\ &= \sum_{p \in m} \sum_{q \in l} \left(\frac{\partial L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)}{\partial W_{pq}^{(k)}} \right)^2 \\ &= \sum_{p \in m} \sum_{q \in l} \left(\frac{\partial L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)}{\partial Z_{ip}^{(k+1)}} H_{iq}^{(k)} \right)^2 \\ &= \left(\sum_{p \in m} \frac{\partial L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)}{\partial Z_{ip}^{(k+1)}} \right)^2 \left(\sum_{q \in l} H_{iq}^{(k)} \right)^2 \end{aligned} \quad (37)$$

Therefore, we just need to compute the product of the square sum of $\frac{\partial L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)}{\partial Z_i^{(k+1)}}$ and $H_i^{(k)}$, which are all from intermediate results during the computation of mini-batch SGD. Its computation complexity is just $O(b(m+l))$, which is extremely light-weight considering that the cost for calculating the gradient is $O(bml)$. Algorithm 4 shows the vectorized computation of Active Sampler in each layer of MLP. For deep models which contains multiple layers, the square of the gradient magnitude with respect to parameters from whole layers can be computed by summing the square of gradient magnitude with respect to parameters from each layer, i.e.

$$\|\nabla_{\mathbf{w}} L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)\|_2^2 = \sum_k \|\nabla_{W^{(k)}} L(f_{\mathbf{w}}(\mathbf{x}_i), y_i)\|_2^2 \quad (38)$$

4. EXPERIMENTAL STUDY

4.1 Experiment Setup

We shall evaluate the speedup of Active Sampler using a set of popular benchmark tasks, namely the hand-written digit classification on MNIST using SVM [25], malicious URL detection in URL using feature selection [16], and image classification on CIFAR10 using CNN [12]. In addition, we shall also test the scalability of Active Sampler using the CIFAR10 dataset with data augmentation, where the size of the training data is increased by 128x. Table 3 summarizes the datasets and models used in our experimental study.

- **MNIST:** MNIST is a benchmark dataset of handwritten digits classification, consisting of a training set of 60000 images and a test set of 10000 images. Each image contains 28×28 gray pixels. Pegasos [25] is a mini-batch SGD solver for kernel SVM model. The test error of kernel SVM in MNIST dataset is 0.60% [25].
- **URL:** URL [18] is a dataset for malicious URL detection. It consists of 2.4 million URLs and 3.2 million features. Each URL contains around 100 non-zero features and hence its features are quite sparse. Lasso regression [16] is a popular feature selection model as described in Table 1. Its test error in the URL dataset is around 2.5%.
- **CIFAR10:** CIFAR10 is a dataset for image classification, consisting of a training set of 60000 images and a test set of 10000 images. It is the benchmark dataset commonly used for the evaluation of deep convolutional neural network (DCNN) [12] models. Each image contains 32×32 colored pixels. Its test error without data augmentation is 18%.
- **CIFAR-DA:** Data augmentation is a standard technique to increase the size of training data. It generates additional images by slightly translating the original images. We use the data augmentation version of the CIFAR10 dataset (CIFAR-DA) to study the scalability of Active Sampler. It contains 128x images compared with CIFAR10. Its test error for DCNN model is 11.5%. However, as the number of training images increases, DCNN model takes significantly longer time to achieve its best performance.

All the models are trained under the SGD framework. The standard mini-batch SGD (**MBSGD**) algorithm is used as the baseline. We also implement the mini-batch version of Active Sampler for comparison, with the same size of mini-batch. Unless otherwise specified, the size of mini-batch is set to 128. The validation accuracy are tested per 100 mini-batch iterations. To study the effect of History Reinforcement strategy described in Section 3.4.1, we have implemented two versions of Active Sampler, with and without History Reinforcement. **ASSGD**, the Active Sampler without History Reinforcement, is expected to perform well for moderate size of training examples, while **ASHR**, the Active Sampler with History Reinforcement, is expected to yield better performance for large-scale training sets. In ASHR, the whole dataset is randomly split into 16 large batches, and examples are trained 16 times on average at each stage.

All of the algorithms are implemented in C++, compiled using GCC O2, and OpenBlas is adopted to accelerate linear algebra operations. Experiments for MNIST, URL and CIFAR10 are carried on an Intel Xeon 24-core server with 500GB memory.

Distributed Environment and Scalability Test:

We also study the performance of Active Sampler in a dis-

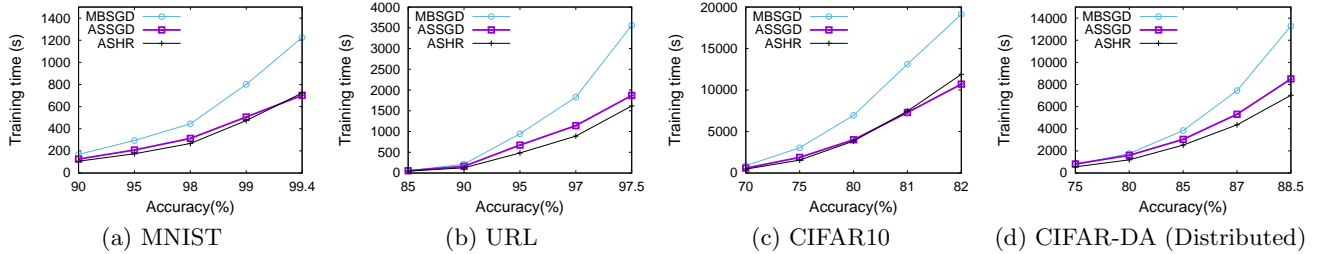


Figure 6: Overall Training Time

distributed environment with the scale of millions of training examples. In general, the benefit of Active Sampler is independent to the architecture of the training system as long as it is still under the SGD framework. SGD training algorithms can be easily distributed to clusters via the parameter server [14] architecture. The main difference between a distributed SGD and a single-node SGD is that the distributed SGD incurs additional communication overhead. Since the communication cost for a mini-batch is a constant while the computation cost is proportional to its size, distributed SGD usually uses a larger mini-batch to reduce the proportion of communication cost. We conduct our scalability study using CIFAR-DA dataset on the Apache SINGA system [28], which is a general distributed deep learning platform. We follow all the default settings of CIFAR on SINGA, where the mini-batch size is set to be 512. The distributed environment is a 32-node cluster, where each machine is equipped with an Intel Xeon 4-core CPU and 8GB memory.

4.2 Overall Performance

Figure 6 shows the training time to reach a certain accuracy for MBSGD, ASSGD and ASHR. 99.4%, 82%, 97.5% and 88.5% are the best accuracy achieved in these four tasks respectively. Generally, the convergence of ASSGD and ASHR are significantly faster than MBSGD. To reach the optimal test error, ASSGD and ASHR save about 40% to 60% of the training time. The speedup is especially great in the latter stages of training, as evidenced by the bigger difference in the slope between the algorithms in the right hand portion of each graph. A possible explanation to this phenomenon is that the models typically have smaller changes near the end stage of training. As a result, the larger variance in MBSGD would have more serious negative effect, while ASSGD and ASHR algorithm would get even better approximation of the scale of gradient. There are also some notable differences between the performance of ASSGD and ASHR. First, ASHR converges much faster than ASSGD in the two large datasets (URL and CIFAR-DA), demonstrating that ASHR provides more accurate approximation of the scale of gradient in large-scale datasets. Meanwhile, its speed-up is less than ASSGD in the two small datasets (MNIST and CIFAR10), probably due to only a subset of training data used by ASHR which only contains around 3000 training examples. Second, ASHR converges slightly faster at the beginning, and slightly slower near the end. This is because ASSGD needs to visit the whole dataset at least once before enjoying the benefit of smaller variance, while ASHR only needs to visit a subset of the dataset. However, in later stages of training, ASSGD gets the gradient approximation as accurate as ASHR since the model

change is not significant, while ASHR is still suffering from the bias introduced by sampling from partial data.

For the scalability test on CIFAR-DA, its training time is even smaller than CIFAR10 due to distributed training. Active Sampler still works as expected: ASHR speeds up the training process by 1.9x, and ASSGD speeds up the process by 1.6x. This is because the benefit of Active Sampler is derived from the total number of iterations used to achieve a certain accuracy, instead of the improvement of training time per iteration. Since the total number of iterations used is not affected by the training architecture, the speed-up of Active Sampler is applicable to all kinds of SGD frameworks as long as its overhead in the training time per iteration is small. Conversely, the number of training examples does affect the performance of ASSGD, since its approximation becomes less accurate when the number of training examples increases. However, ASHR scales well in all cases.

4.3 Variance of Stochastic Gradients

From the stochastic optimization view point, the benefit of using Active Sampler is derived mainly from the reduction of the variance of stochastic gradient. We therefore evaluate the average variance of MBSGD, ASSGD and ASHR and summarize the results in Figure 7. Since the absolute value of variance may change dramatically during the training process, we use the variance of MBSGD as the baseline and report the relative ratio of the variance of ASSGD and ASHR compared with the baseline. The results show that ASHR has the smallest variance, less than 40% of the variance of MBSGD on average. The variance of ASSGD is slightly higher than that of ASHR, especially in the two large datasets, URL and CIFAR10, mainly because that its gradient approximation is less accurate in larger datasets. However, the variance of ASSGD is still less than half of the variance of MBSGD with the same mini-batch size. Another observable trend is that the variance ratio of ASSGD and ASHR are getting smaller with the increase of training time, suggesting that the history gradient approximation is getting more and more accurate as the training time increases.

We note that in MBSGD, the variance is proportional to the reverse of the size of mini-batch. Therefore, to get the same level of variance in the stochastic gradient as Active Sampler, MBSGD needs to increase its mini-batch size by 2-3x. From this angle, Active Sampler is a much more efficient method to reduce the variance of stochastic gradient, instead of relying on the use of a larger mini-batch.

4.4 Training Time Analysis

The overall training time is determined by the product of the training time per iteration and the number of iterations

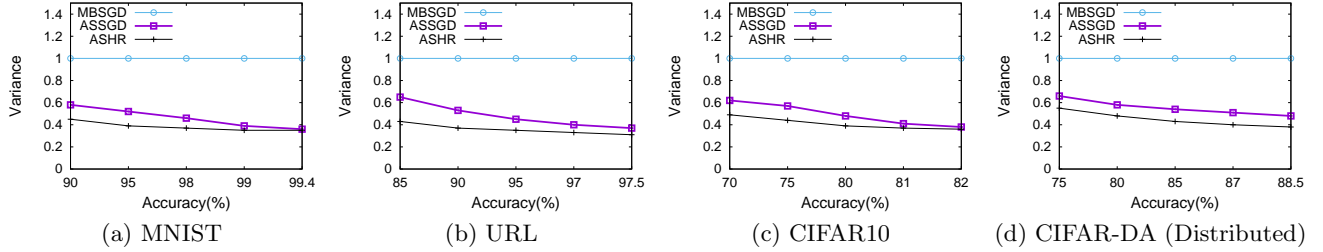


Figure 7: Variance at Different Training Stages

Table 4: Training Time per Iteration

Dataset	MBSGD	ASSGD	ASHR
MNIST	0.179s	0.208s	0.205s
URL	0.080s	0.092s	0.092s
CIFAR10	0.245s	0.295s	0.296s
CIFAR-DA	0.110s	0.119s	0.119s

to reach a certain accuracy. Here we present a detailed study of how Active Sampler affects these two aspects.

Table 4 shows the training time per iteration of MBSGD, ASSGD and ASHR. Obviously, MBSGD is the fastest since ASSGD and ASHR entail additional computations. However, the difference is not significant. ASSGD and ASHR only require 15%-20% more time than MBSGD, while providing the stochastic gradient with much smaller variance. As discussed above, to reach the same variance, MBSGD needs to use 2-3x samples in a mini-batch, which may incur 100%-200% additional overhead. In the distributed trained task CIFAR-DA, the overhead of ASSGD and ASHR are even smaller, which are around 10%. This is because ASSGD and ASHR do not incur any overhead to the communication costs. There are no major differences between ASSGD and ASHR, since they have exactly the same computation logics inside each iteration.

Figure 8 shows the number of iterations to reach a certain accuracy. The number of iterations required by ASSGD and ASHR is around 40% to 60% of the number of iterations required by MBSGD. The proportion of iterations saved varies with different datasets. The iterations saved would be more significant when the contribution from training examples are highly biased. However, theoretically, it is possible that all examples have the similar effect on refining the model (an extreme case is all examples being the same), and uniform sampling becomes the optimal weighted sampling. Not surprisingly, as indicated by the experiments so far, all of the benchmark datasets do not represent the extreme case, and a significant number of training iterations can be saved.

5. RELATED WORK

Complex machine learning models, such as large-scale linear methods[25], feature selection [16] or deep learning [4], are widely adopted in Big Data analytics. Due to the huge size of both model and data, how to train these model efficiently is a challenging topic, and the solution requires efforts from learning, database, and system communities. Many optimizations have been proposed from a systems perspective for specific classes of models [31, 32, 13, 30, 4, 7]. Most of these algorithms (and many others) can fit into

an Empirical Risk Minimization [26] (ERM) framework, for which we aim to develop a more general accelerator.

The optimization of the general ERM is widely studied in machine learning community [26]. Generally, there are two classes of methods: first-order algorithms such as gradient descent [1], and second-order algorithms such as Newton method [5]. Although second-order algorithms typically have a much faster convergence rate, they require the Hessian matrix [3] of parameters, making them not practical for large-scale models where the number of parameter is huge. For similar reasons, batch gradient methods [29] are very expensive for large training datasets. Therefore, stochastic methods [22] are the most favored algorithm in recent large-scale machine learning applications.

Stochastic Gradient Descent [22] (SGD) is one of the most popular stochastic optimization methods. Theoretical results are well studied in [20]. However, [27] has shown that the variance in stochastic gradient is the key factor limiting the convergence rate of SGD. Consequently, many SGD variants such as SAG [23], SVRG [11], S3DG [19] have been developed to reduce the variance. The convergence rate of these variants has been greatly improved in both theory and practice in terms of the number of iterations required to reach a certain accuracy. However, the optimization cost of these methods are not negligible, causing the training cost per iteration to increase substantially.

There are also studies [6] on the effect of learning rate on the convergence rate of SGD. Naturally, reducing the multiplier of gradient in updates will reduce the variance in each update. This idea motivates us to study if we can scale down those stochastic gradients with larger variance by using a smaller learning rate, while making up the effects of those gradients by increasing their sampling frequency. Based on this intuition, we propose to accelerate the SGD training based on the idea of active learning [24, 21]. Active learning was originally proposed to select a set of labeled training data to maximize the accuracy of model. [17] uses the idea of weighted sampling to maximize the information gain of active learning. However, in our Active Sampler, all training data are already labeled, and the active selection is to maximize the learning speed of a passive learning model.

Active Sampler is also related to feature selection methods [31]. Both of them assume that not all the training data are informative for model construction. The difference is that feature selection methods find the most informative columns in the training data, whereas Active Sampler finds the most informative rows.

6. CONCLUSION

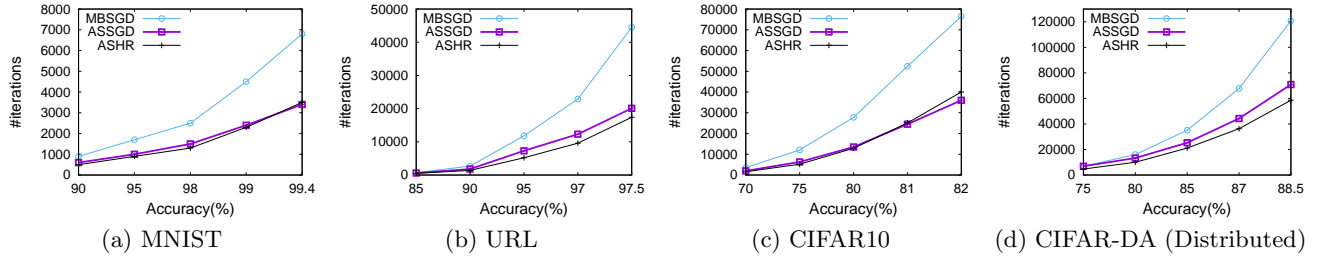


Figure 8: Number of Iterations with respect to Accuracy

SGD algorithms are playing a central role in the model training of complex data analytics, where sampled training data are used at each training iteration. Uniform sampling and sequential access have been commonly used due to their simplicity. In this paper, we study how the sampling method can affect the training speed as a means to facilitate analytics at scale. Based on the inspiration from active learning, we propose Active Sampler which has sampling frequency that is proportional to the magnitude of gradient. We show the correctness and optimality of Active Sampler in theory, and developed a set of schemes to make the implementation light-weight. Experiments show that Active Sampler can speedup the training procedure of SVM, feature selection and deep learning by 1.6-2.2x, compared with the uniform sampling. Results also demonstrate that Active Sampler has a significant effect on reducing the variance of the stochastic gradient, making the training process much more stable.

7. REFERENCES

- [1] D. P. Bertsekas. Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. *Optimization for Machine Learning*, 2011.
- [2] L. Bottou. Large-scale machine learning with stochastic gradient descent. In *COMPSTAT*, 2010.
- [3] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.
- [4] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, A. Senior, P. Tucker, K. Yang, Q. V. Le, et al. Large scale distributed deep networks. In *NIPS*, 2012.
- [5] J. E. Dennis, Jr and J. J. Moré. Quasi-newton methods, motivation and theory. *SIAM Review*.
- [6] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *JMLR*, 2011.
- [7] T. Elgamal, M. Yabandeh, A. Aboulnaga, W. Mustafa, and M. Hefeeda. spca: Scalable principal component analysis for big data on distributed platforms. In *SIGMOD*, 2015.
- [8] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *JMLR*, 2013.
- [9] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.
- [10] A. Ilin and T. Raiko. Practical approaches to principal component analysis in the presence of missing values. *JMLR*, 2010.
- [11] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *NIPS*, 2013.
- [12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.
- [13] A. Kumar, J. Naughton, and J. M. Patel. Learning generalized linear models over normalized data. In *SIGMOD*, 2015.
- [14] M. Li, D. G. Andersen, J. W. Park, A. J. Smola, A. Ahmed, V. Josifovski, J. Long, E. J. Shekita, and B.-Y. Su. Scaling distributed machine learning with the parameter server. In *OSDI*, 2014.
- [15] M. Li, T. Zhang, Y. Chen, and A. J. Smola. Efficient mini-batch training for stochastic optimization. In *SIGKDD*, 2014.
- [16] J. Liu, J. Chen, and J. Ye. Large-scale sparse logistic regression. In *SIGKDD*, 2009.
- [17] B. Long, O. Chapelle, Y. Zhang, Y. Chang, Z. Zheng, and B. Tseng. Active learning for ranking through expected loss optimization. In *SIGIR*, 2010.
- [18] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker. Identifying suspicious urls: an application of large-scale online learning. In *ICML*, 2009.
- [19] Y. Mu, W. Liu, and W. Fan. Stochastic gradient made stable: A manifold propagation approach for large-scale optimization. *arXiv:1506.08350*, 2015.
- [20] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 2009.
- [21] M. Prince. Does active learning work. a review of the research. *Journal of Engineering Education-Washington*, 2004.
- [22] H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 1951.
- [23] M. Schmidt, N. L. Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. *arXiv:1309.2388*, 2013.
- [24] B. Settles. Active learning literature survey. *University of Wisconsin, Madison*, 2010.
- [25] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter. Pegasos: Primal estimated sub-gradient solver for svm. *Mathematical Programming*, 2011.
- [26] V. N. Vapnik and V. Vapnik. *Statistical learning theory*, volume 1. Wiley New York, 1998.
- [27] C. Wang, X. Chen, A. J. Smola, and E. P. Xing. Variance reduction for stochastic gradient optimization. In *NIPS*, 2013.
- [28] W. Wang, G. Chen, A. T. T. Dinh, J. Gao, B. C. Ooi, K.-L. Tan, and S. Wang. Singa: Putting deep learning in the hands of multimedia users. In *ACM MM*, 2015.
- [29] D. R. Wilson and T. R. Martinez. The general inefficiency of batch training for gradient descent learning. *Neural Networks*, 2003.
- [30] F. Yan, O. Ruwase, Y. He, and T. Chilimbi. Performance modeling and scalability optimization of distributed deep learning systems. In *SIGKDD*, 2015.
- [31] C. Zhang, A. Kumar, and C. Ré. Materialization optimizations for feature selection workloads. In *SIGMOD*, 2014.
- [32] C. Zhang and C. Ré. Towards high-throughput gibbs sampling at scale: A study across storage managers. In *SIGMOD*, 2013.